Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 54 (2015) 390 – 395

Procedia Computer Science

Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Finding Similar Patents Through Semantic Query Expansion

Pawan Sharma[a,*], Rashmi Tripathi[b] and R. C. Tripathi[c]

[a] *Indian Institute of Information Technology, Allahabad 211 002, U.P, India*
[b] *Indian Institute of Information Technology, Allahabad 211 006, U.P, India*
[c] *Indian Institute of Information Technology, Allahabad 211 012, U.P, India*

Abstract

Patent search is a complex task and involves a great level of expertise. Through this research we have tried to find similar patents by expanding the user query semantically. The main purpose of this research is to investigate how the patent retrieval system can be improved by using words which have same expression i.e. semantically similar. WorldNet and Wikipedia are used as an external source for expanding the query. Result shows that expanded query yields better results compared to conventional approaches of patent search.

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Keywords: Patent search; Query expansion; Retrieval; Semantic similarity; Wordnet; Wikipedia.

1. Introduction

Patents documents provide an important source for keeping a watch on past and ongoing innovation. Patent is granted for new/novel device equipment method etc. having a definite step of invention. Before filing a patent, inventor gets a rigorous prior art search performed through a reliable source. Patent search is a complex task and requires high skill ability[1]. With the advent of new inventions growing need of patent searches, and new patent search tools is being realized. These tools work mostly using techniques of text mining and keyword search for identifying almost similar patents of past. Multiple keywords and their synonyms are used numerous times manually to find the similar or so called similar patent as a exercise of prior art search. However the drawback with these approaches is that they depend on writing style of authors and selection of words consuming ample time. The search process of automated tools uses exact key word or Bag of Words (BoW)[2] approach. BoW has its own limitations. It can fetch out patents with exact similar words without any regard of their meaning in the required contest. Most words and their phrases have different meanings in different Contests. In such a situation it is almost impossible to find out all the synonyms of a word and find similar patents. For example the word '*train*' is sometimes used as phrase '*train dataset*' and '*train route*'. Prior one is being used in computational aspects where as the later as a means of transport, thus representing two completely different domains.

Semantic solutions are considered as a good measure for solution above such problems. It improves the recall rates with least effect on precision. In the current research, we have tried to apply semantic expansion technique

*Corresponding author.
E-mail address: rs102@iiita.ac.in

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)
doi:10.1016/j.procs.2015.06.045

to enhance the search process. The main purpose of the current research is to investigate how the patent retrieval system can be improved by using words which have same expression i.e. semantically similar. Semantic expansion methods are a better technique comparatively as it incorporates external sources for expansion. Word Net[3] and Wikipedia are used as external source in current research. Further we have calculated similarity using Cosine[4] and Extended Jaccard coefficient[5] and compared the results. The purpose of similarity testing is to find the best similar results.

2. Motivation

Patent search is a complex task and employs a lot of manpower. Search process consumes ample time resulting delay in granting of a patent. If an automated tool can be developed to ease the process of patent search, the grant procedure will be more responsive and quick. The hypothesis explored in this research is that by using semantic expansion of a query, we can provide high quality results and reduce the complex search process. Expansion technique semantically increases the number of words in a query which gives a better quality of retrieval result. The longer the query abstract, the higher the result accuracy rate.

3. Related Work

Semantic similarity[6] has been used previously by researchers for identifying the relation between two key words. Cui *et al.*[7] has used semantic expansion of query to find solution for query which is well defined in a text. Cui used query log for query and document term and find a solution of probabilistic correlation. Large size of query log helps in improved retrieval results. Wong *et al.*[8] used Query expansion and identified phrases from the query. These phrases are used to identify similarity phrases from database. The identified phrases are weighed with help of various algorithms. Subramanian[9] suggested use of improved Stemming Algorithm for data pre processing to save both space and time and use of links analysis techniques in information retrieval. The incoming and outgoing links are measured for expansion of query. Metzler *et al.*[10] examines a range of similarity measures, including purely lexical measures, stemming, and language modelling – based measures.

4. Methodology

The basic steps involved in our computation of similarity are,

(i) Filtering the abstracts from patent databases on the basis of International Patent Classification (IPC).
(ii) Process abstracts to generate keyword vector.
(iii) Expand Query and construct vector space.
(iv) Finally calculate Similarities.

Patent abstract is used as a query by user. User input abstract as a query and (IPC) as a metadata. Use of IPC as a metadata reduces the search time and helps to focus based on classification of patents in the specified domain. IPC is a classification system which divided each innovation according to its domain e.g. Life science, Technology etc. Patent Abstract may vary from 20 to 400 words or more.

4.1 Indexing

Abstract from the metadata class are indexed[11]. Without using an index, retrieval systems will process through the entire abstract to search for similar kinds. Hash – table files are used for indexing data structure. With Hash – table it is easier to calculate Term Frequency – Inverse Document Frequency (TF-IDF)[12] which is used to determine the weights of index terms inside a document vector. TF-IDF is used to distinguish between relevancy and non-relevant abstracts based on appearance of a term in the abstract text.

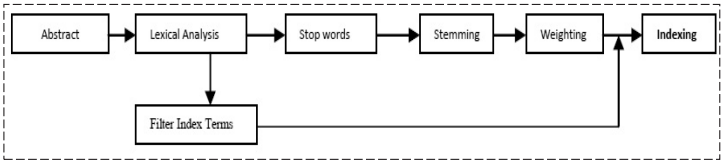

Fig. 1. Pre-processing of an abstract query.

4.2 Indexing steps

Lexical analysis[13] is the primary step of indexing which performs the task of streaming text to stream of index terms. Word which is used to form a sentence is not so useful in retrieval. Such words like a, an, is, are termed as Stop Words. Deleting these stop words is an important step before indexing system. It reduces the size of indexing system and processing time of retrieval. A list of stop words used in patents is made available from European Patent Office official website. *Stemming*[14] is another step of transforming a word to its root or stem and tokens are extracted from the query abstract. The final step of indexing deals with term weighting. Index terms are weighted differently according to their significance in an abstract. Such weighting can be binary. The following code is used in our system to tokenize all abstract from the database. We ignore spaces, punctuations and digits.

$$\text{tPunct}|\text{tSpace}|(\text{tAlpha} * \text{tDigit} + \text{tAlpha}*)$$

Figure 1 sketches the preprocessing phase performed in the system taking an abstract as a query and yielding its index terms.

4.3 Query expansion

The novelty of this Research is query expansion using an external knowledge source. The external knowledge source applied in current research is Word Net as one of the basis to find the meaning of the word and Wikipedia as another knowledge foundation, which tells us about the relations between the given word and other words lexically non similar with the query word. Patent document has many technical terms and Wikipedia source is the best available option which can be used to understand the meaning of these terms. Recall rate has more importance in patent retrieval system whereas other retrieval system emphasis precision – oriented[15]. Recall rates can be improved by query expansion.

4.4 Query expansion with word net

JAWS library are used for implements the (is-a) relation of WorldNet. IS-A[16] relation helps to expand each word from query words characteristic same in nature. Hypernym – hyponymy relations, and noun word of a common concept is used for expansion[17]. Nouns have a significant importance in a sentence. Dissimilar weights are assigned to each token for expanding in a vector[18]. Extended tokens are non similar in meaning then the original word. We add the given word in the vector with a weight equals to 1, and then all the other expanded tokens have a weight in (0,1). Each word weight and depth is measured, for example for word ‘speech’, the words ‘talk’ and ‘lecture’ are in depth 1 and words ‘language’ and ‘debate’ are in depth 2. Assign a higher weight to the tokens closer to the given word and lower weight to the further ones. Weight is calculated by given link’s weight to the power of the token’s depth (distance from original token). Weight and Depth calculation formula is shown (1) below, where W refers to weight and D refers to depth. Token is shown as T and Link is marked as L.

$$^{W}T = {}^{W}L^{D} \tag{1}$$

4.5 Query expansion with wikipedia

To choose more reliable page titles for expansion, we choose links which are more important than the others. Link structure is used for query expansion through Wikipedia. Each word is provided a Hyperlink which is connected with

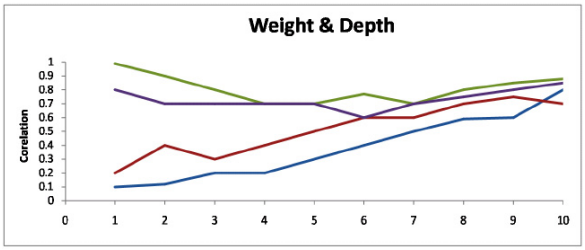

Fig. 2. Effect of weight and depths for Hypernym and hyponym for word net and wikipedia

a page. Expand the query by page title of incoming and outgoing links. Page Rank[19] algorithms suggest numbers of incoming links are measured as an important degree and are more reliable for expansions. Incoming links have a greater importance than outgoing links. Both incoming and outgoing links are involved for effectiveness. To construct an incoming vector for each token we need Incoming Links (Depth and Weight)[20], Threshold and Expansion Size (number of links). Distances from the original page are considered as link weight and the distance as depth. Outgoing vector is constructed in the same way as an incoming vector; however instead of incoming Links Depth and Incoming Links Weights, we use Outgoing Links Weights and Outgoing Links Depth. Each Wikipedia web page has links ranging from 10 to more than 100. Limits on expansion size i.e. number of links have to be determined to avoid noise and failure. The expansion is uniform nature as it will restrict the size of the vectors. Limiting the expansion size shows the number of expanded tokens per each token. The effects of Weights and Depths values in semantic expansion have to be calculated. The Fig. 2 graph shows Hypernym and Hyponym for weight and depth values for WorldNet and Wikipedia. Graph line in green and purple are Hyponymy for WorldNet and Wikipedia while lines in blue and red are Hypernymy for WorldNet and Wikipedia.

4.6 Similarity finding

The main purpose of semantic – based similarity measures is using knowledge sources to measure similarity based on the content of the words rather than frequency of occurrence. Expanded vectors[20] are used to measure similarity between two units of language using cosine similarity and extended Jaccard Coefficient. In Cosine similarity method abstract are represents as a multi dimensional vector and each token in the vector is represented as one dimension. Extended Jaccard is an advancement of Jaccard. Jaccard can only be used for binary vectors i.e. similarity between objects of purely binary attributes Extended Jaccard similarity retains the thinness property of the cosine while admitting unfairness of collinear vectors. Formula for Ext. Jaccard coefficient is shown in (2).

$$\text{Extended Jaccard Similarity } (v_1, v_2) = \frac{v_1 \times v_2}{v_1^2 + v_2^2 - v_1 \times v_2} \tag{2}$$

5. Implementation and Results

In order to verify the validity of model, the author conducts two experiments. The first experiment is conducted without query expansions while the other after query expansion. Both the results are compared in terms of Precision and recall rates. Recall and precision are value of single metrics. For retrieval systems which return ranked list results, it is better to take average precision and mean average precision method into account, for measuring the performance of a system. Recall and precision cannot determine it. In average precision the relevant documents with better rank gain more weights and it is calculated as in formula 3:

$$\text{Average Precision} = \frac{\sum_{r=1}^{N} (P(r) \times \text{Rel}(r))}{(|\{\text{Relevant documents}\}|)} \tag{3}$$

Table 1. User's query compared result with precision and recall.

Method	Mean Recall	Mean Precision	Average Precision	Mean Avr. Precision
Without Query Expansion	81.1%	5.07%	6.01%	25.11%
With Query expansion	94.40%	5.13%	6.36%	16.32%

Table 2. Similarity results using cosine similarity method.

Patent→ query↓	P1	P2	P3	P4	P5	P6	P7	P8	P9
Q1	.981	.980	.891	.800	.871	.871	.871	.871	.871
Q2	.910	.863	.771	.700	.671	.603	.590	.550	.541
Q3	.881	.894	.831	.702	.681	.610	.608	.532	.512

Table 3. Similarity results through Ext. Jaccard coefficient.

Patent→ query↓	P1	P2	P3	P4	P5	P6	P7	P8	P9
Q1	−0.601	0.721	−0.734	0.843	0.945	−1.321	−1.336	−1.575	−1.594
Q2	−0.201	−0.331	−0.384	0.445	0.654	−0.882	−1.243	−1.446	−1.413
Q3	−0.067	−0.219	0.346	0.575	0.685	−1.011	−1.226	−1.355	−1.450

Mean Average Precision (MAP) is a metric for calculating the mean of average precision per query. It measures results on the basis of a set of queries and is measured as in formula 4:

$$\text{MAP} = \frac{\sum_{q=1}^{N} \text{AveP}(q)}{(|\{Q\}|)} \tag{4}$$

We also calculated mean recall and mean precision according to the formula (5) (6).

$$\text{Mean Recall} = \frac{\sum_{q=1}^{N} \text{Recall}(q)}{(|\{q\}|)} \tag{5}$$

$$\text{Mean Precision} = \frac{\sum_{q=1}^{N} \text{Precision}(q)}{(|\{Q\}|)} \tag{6}$$

Table 1 shows the comparison results of precision, recall, average precision and mean average precision obtained for a search with and without query expansion for a single query. Expanded Query shows more improved results than Query without expansion.

Measure of similarity is done through Extended Jaccard coefficient and Cosine similarity. We asked an expert of computer science field to validate our results based on his knowledge. Table 2 shows result of cosine similarity and expert view on the results. The numbers in bold are expert opinion regarding similarity between the query and system identified similarity.

Table 3 shows result of Ext. Jaccard coefficient and expert view on the results. The numbers in bold are expert opinion regarding similarity between the query and system identified similarity.

Our experts regarded the bold marked patents in Table 2 and 3, semantically similar. As the result of comparison among two measurements, the cosine similarity was the closest to the judgments by our experts and the one of Ex. Jaccard was weak in identifying similarity. The normalized semantic similarity by Ex. Jaccard coefficient was not found to correlate with the judgment by our experts.

6. Discussion

The reason for Ex. Jaccard coefficient weakness is that it does not consider term frequencies. In the formulation of Ex. Jaccard coefficient, technical terms which specify the topic and common terms, which appear in many documents,

have the same weight (significance). These results indicate that since significant terms tend to appear several times, term frequency should be considered as significant to measure the semantic similarities. Our result indicates that comparing the cosine similarities of TF-IDF vectors between patents enables us to obtain similar patents. This proposed approach can not only provide a new viewpoint in identifying similar patents but can also reduce the human labor involved in patent searching.

7. Conclusion

In this paper, an attempt was made to identify similar patents. This research used Indian Patent database[21] in excel form. Authors expanded the query by using external source like Wikipedia and WorldNet to find the meaning and relation among two words. The enhanced query results into better results compared with non expanded query in terms of Recall rates. We further found that cosine similarity method results into better similarity finding rather than Ex. Jaccard coefficient. A case study was performed for an ICT related patents and expert opinion was taken into consideration. As a result, the cosine similarity was found the best way to discover the corresponding similar patents. This proposed approach can be worked further as a measure for prior art search.

References

[1] H. Mase, T. Matsubayashi, Y. Ogawa, M. Iwayama and T. Oshio, Proposal of Two – Stage Patent Retrieval Method Considering the Claim Structure, vol. 4, pp. 190–206, June (2005). [Online]. Available: http: //doi.acm.org/10.1145/1105696.1105702.

[2] H. Lashkari, F. Mahdavi and V. Ghomi, A Boolean Model in Information Retrieval for Search Engines, *International Conference on Information Management and Engineering*, pp. 385–389, (2009).

[3] J. Kamps and M. Koolen, Is Wikipedia Link Structure Different? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining, Ser. WSDM'09*, New York, NY, USA: ACM, pp. 232–241, (2009).

[4] J. Jiang and D. Conrath, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, In *Proceedings of the International Conference on Research in Computational Linguistics*, Taiwan, (2010).

[5] Y. Kim, J. Ryu and S.-H. Myaeng, A Patent Retrieval Method Using Semantic Annotations, In *KDIR'09*, pp. 211–218, (2009).

[6] Angelos Hliaoutakis, Kaliope Zervanou and Euripides G. M. Petrakis, The AMTEx Approach in the Medical Document Indexing and Retrieval Application, *Data and Knowledge Eng.*, pp. 380–392, (2009).

[7] Hang Cui, Min-Yen Kan and Tat-Seng Chua, Soft Pattern Matching Models for Def Initial Question Answering, *ACM Transactions on Information Systems*, vol. 25(2), April (2007).

[8] S. K. M. Wong, W. Ziarko and P. C. N. Wong, Generalized Vector Space Model in Information Retrieval, In the *8th Annual International ACM IGIR Conference on Research and Development in Information Retrieval*, New York, pp. 18–25, (1985).

[9] C. Ramasubramanian and R. Ramya, Effective Pre-Processing Activities in Text Mining Using Improved Porter's Stemming Algorithm, *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, Issue 12, (December 2013).

[10] Donald Metzler, Rosie Jones, Fuchun Peng and Ruiqiang Zhang, *Improving Search Relevance for Implicitly Temporal Queries*, SIGIR, pp. 700–701, (2009).

[11] Daniel Bar, Chris Biemann, Iryna Gurevych and Torsten Zesch, UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures, First Joint Conference on Lexical and Computational Semantics (*SEM), pp. 435–440, Montreal, Canada, June 7–8, 2012 Association for Computational Linguistics, (2012).

[12] S. Bashir and A. Rauber, Analyzing Document Retrievability in Patent Retrieval Settings, In *Proceedings of the 20th International Conference on Database and Expert Systems Applications, ser. DEXA'09*, Berlin, Heidelberg: Springer-Verlag, pp. 753–760, (2009).

[13] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross and K. Miller, Word Net: An Online Lexical Database, *Int. J. Lexicograph*, vol. 3,4, pp. 235–244, (1990).

[14] R. Resnik, Using Information Content to Evaluate Semantic Similarity, In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, (1995).

[15] C. Leacock and M. Chodorow, Combining Local Context and Word Net Sense Similarity for Word Sense Identification, In *WordNet, An Electronic Lexical Database*, The MIT Press, (1998).

[16] Y. Xu, Apply Text Mining in Analysis of Patent Document, In *IEEE 10th International Conference on Computer-Aided Industrial Design Conceptual Design*, 2009, CAID CD 2009, pp. 2350–2352, November (2009).

[17] K. S. Jones, Idf Term Weighting and IR Research Lessons, *Journal of Documentation*, vol. 60, pp. 521–523, (2004).

[18] V. Patwardhan, Vector-Based Semantic Expansion Approach: An Application to Patent Retrieval, Master's Thesis, Delft University of Technology Delft, The Netherlands, (2011).

[19] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval, New York, NY, USA: McGraw-Hill, *Inc.*, (1986).

[20] G. Salton, A. Wong and C. S. Yang, A Vector Space Model for Automatic Indexing, *Commun. ACM*, vol. 18, pp. 613–620, November (1975).

[21] Pawan Sharma and R. C. Tripathi, *International Journal of Database Management*, vol. 5, no. 5, pp. 9–16, (2013).